Deep Reinforcement Learning for Mobile 5G and Beyond: Fundamentals, Applications and Challenges

Zehui Xiong, Yang Zhang, *Member, IEEE*, Dusit Niyato, *Fellow, IEEE*, Ruilong Deng, *Member, IEEE*, Ping Wang, *Senior Member, IEEE*, Li-Chun Wang, *Fellow, IEEE*

Abstract

Future generation wireless networks, i.e., 5G and beyond, have to accommodate the surging growth of mobile data traffic and to support a high density of mobile users with a variety of services and applications. Meanwhile, the networks become increasingly dense, heterogeneous, decentralized, and ad hoc in nature, involving numerous and diverse network entities. As such, different objectives, e.g., high throughput and low latency, need to be achieved in the service and resource allocation has to be designed and optimized accordingly. However, with the dynamics and uncertainty inherently existing in the wireless network environments, conventional approaches of service and resource management that require complete and perfect knowledge of the systems become inefficient or even inapplicable. Inspired by the success of machine learning in solving complicated control and decision-making problems, in this article, we focus on deep reinforcement learning based approaches, which allow network entities to learn and build knowledge about the networks to make optimal decisions locally and independently. We first present an overview and fundamental concepts of deep reinforcement learning. Next, we review some related works that capitalize deep reinforcement learning to address different issues in 5G networks. Finally, we present an application of deep reinforcement learning in 5G network slicing optimization. The numerical results demonstrate that the proposed approach achieves superior performance compared with baseline solutions.

Index Terms

Mobile 5G, deep reinforcement learning, intelligent resource management, slicing.

Z. Xiong, D. Niyato and R. Deng are with School of Computer Science and Engineering, Nanyang Technological University, Singapore. Y. Zhang is with School of Computer Science and Technology, Wuhan University of Technology, China. P. Wang is with the Department of Electrical Engineering and Computer Science, York University, Canada. L-C. Wang is with Department of Electrical Engineering, National Chiao Tung University, Taiwan.

I. Introduction: The Challenge of Dynamics in 5G and Beyond

The rapid and tremendous increase of wireless data services driven by the popularity of smart mobile devices and communications technologies has triggered the investigation of the future generation wireless networks, i.e., 5G and beyond. 5G networks are expected to support a variety of applications with diverse requirements, including higher peak and user data rates, reduced latency, enhanced system capacity, improved energy efficiency and so on. To achieve this goal, a series of emerging wireless technologies have been proposed, such as massive MIMO (multiple-input-multiple-output) and millimeter-wave (mmWave) communications. With massive MIMO, each base station (BS) can transmit high-speed data streams to multiple user equipments (UEs) simultaneously. Alternatively, with mmWave communications, the BS can efficiently exploit high-frequency spectrum for overcoming bandwidth shortage and providing UEs with more available bandwidth.

Apart from physical network advancements, i.e., massive MIMO and mmWave, another major drive in 5G and beyond is the network softwarization. It can provide more flexibility for mobile service management under dynamic network conditions and service demands. Meanwhile, with the above techniques supporting a large number of UEs, the 5G network becomes increasingly heterogeneous and decentralized in nature. Multiple network entities are also involved, e.g., BSs with different types and UEs with different QoS requirements. Optimized decisions of the network entities need to be determined towards different objectives, such as data rate maximization, network latency, and energy consumption minimization. However, it is challenging to achieve the optimized resource and service management in the presence of the wide variety of service requirements as well as the inherent dynamics and uncertainty in mobile 5G network environments, such as fast time-varying wireless channels and constantly changing network topology [1].

In such time-varying and unpredictable network environments, Reinforcement Learning (RL) has shown to be a viable tool to tackle the real-time dynamic decision-making problems [2]. On the one hand, instead of myopically optimizing the current benefits, RL naturally incorporates farsighted system evolution when making decisions, which is essential for time-variant 5G networks. On the other hand, RL can update decision policies to reach optimal system performance through the reward feedback of the previous decisions, i.e., reinforcement, even without up-to-date information [2]. Therefore, RL-based approaches can be an option for solving resource and

service management problems in mobile networks. Nevertheless, conventional RL algorithms, such as Q Learning, suffer from slow convergence speed, especially if the state space and action space of the problem are large. Furthermore, the algorithms have to store full tables of an immediate value, e.g., Q-value, for each state-action pair. The tables can be too large to be maintained in mobile devices. In this regard, the RL often leads to poor performance. As a branch of artificial intelligence, Deep Reinforcement Learning (DRL), which is the combination of RL and deep learning, has been promisingly proposed to overcome the limitations of RL. DRL has shown vast successes in many applications such as natural language processing and robotics [3].

As DRL possesses great potential in handling large-scale and dynamic systems, we explore the DRL-based approach for mobile 5G and beyond to enhance network performance. In DRL, a deep neural network called Deep Q Network (DQN) is utilized to accelerate the learning process and also to reduce the memory required to store the parameters of the model, making it perfectly suitable for mobile devices with limited resources. As a result, employing DRL is thus promising to address mobile service and network management and control problems under complex, dense, and heterogeneous mobile 5G environments.

In this article, we first present the basic concepts of DRL, and then review some recent works of applying DRL to address problems arising in mobile 5G and beyond, such as power control, offloading and edge caching. Next, we introduce the DRL-based scheme for network slicing optimization, and show that the proposed scheme achieves the best performance compared with other approaches. Finally, we conclude our work and also outline important future research directions.

II. Overview of Deep Reinforcement Learning

In this section, we present the fundamental of reinforcement learning and then discuss the motivation to evolve from reinforcement learning to deep reinforcement learning.

A. Reinforcement Learning

As one of the important machine learning techniques, Reinforcement Learning (RL) enables the optimization of decision making by an agent without a priori knowledge of the system and environment. The agent takes actions at a certain system state and observes the corresponding responses from the environment. The agent either receives a reward for taking the good action

or a penalty for taking the bad action. The agent then adopts a trial-and-error search for possible optimal state-action pairs, referred to as a policy, when making the decision. The agent is encouraged to take a series of actions to maximize the long-term reward of the agent based on past knowledge, namely, reinforcement.

Q Learning algorithm is amongst the most well-known model-free RL algorithms for computing an optimal policy that maximizes the long-term reward. A reward function Q is introduced to map a state-action pair to the expected cumulative reward (i.e., Q-value) for the agent to estimate and decide optimal actions in response to different system states. By recording all the actions which maximize the Q-value, the agent obtains a list of optimal state-action pairs, defined as the optimal policy.

Reinforcement learning is a promising tool for solving many resource management and other optimization issues in mobile communication systems with temporal variation and stochasticity of service and resource availability, as well as system parameters and states. To optimize system services and resources, the agent of a network entity calculates and updates the Q-function for the optimal actions. With the uncertainties in mobile communication systems, as well as the short-sighted perspective of the agent, the rewards gained by the agent can be different when the same actions are taken at the same system states. In this case, the Q-function needs to be updated iteratively until convergence with a properly set updating rate, i.e., learning rate. The Q-function updating process is defined as training.

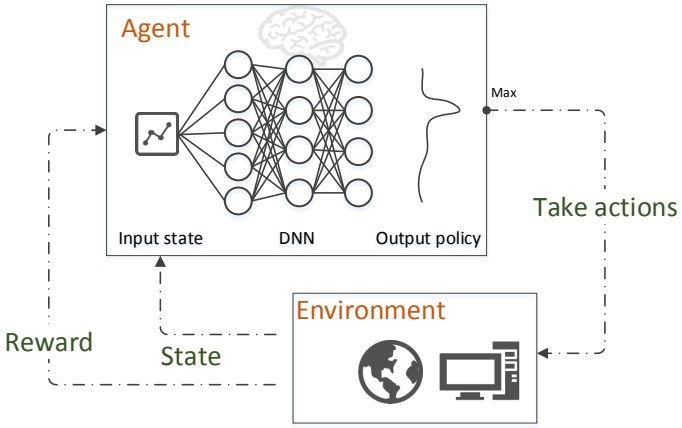


Fig. 1: The illustration of deep Q-learning.

B. From Reinforcement Learning to Deep Reinforcement Learning

Q Learning requires the agent to maintain and update a set of Q-values of all the state-action pairs. However, the future generation wireless networks are expected to be large-scale,

highly heterogeneous and decentralized. Consequently, the number of possible values of system states becomes intractable large. Moreover, with the high diversity and uncertainty of the system components and environment parameters, there can be hidden system states or even an unlimited number of system states. In this case, the cost of calculating and maintaining all Q-values becomes practically infeasible. This situation is referred to as the curse of dimensionality.

To address this issue, Deep Reinforcement Learning (DRL) combines RL with deep learning techniques. Deep Q Learning, which employs a Deep Q Network (DQN), is one of the typical DRL models that apply a deep neural network as an approximation of the Q-function (See Figure 1). Note that the reward and state in Figure 1 can be defined based on different system objective in 5G, such as throughput maximization or outage probability minimization, and power consumption or channel gains. Similar to the Q-function in RL, the DQN takes current observed system states as the input. All the input states are transferred to different neural network layers with some particular weight factors. Finally, the DQN outputs a set of Q-values with respect to all the possible actions. The goal of learning in the DQN is to train and find the most feasible weight factors from historical data, i.e., historical Q-values, actions, and state transitions. For a DQN with multilayer perceptron as the underlying neural network, the complexity of calculating the Q-values and actions is linear. Moreover, the number of inputs of the DQN is only decided by the type of all the states. For each input, various values of the state can be transferred in, e.g., different channel states, without changing the structure of the DQN, even when the number of all the values reaches infinity. If the Q-function is constructed by the neural network, the major differences between RL and DRL lie in the following two aspects: 1) In the RL, the arriving samples are used to train neural network's parameters, while the DRL randomly selects the samples from memory pool to train the DQN's parameters. 2) DRL updates the weight every several time steps to reduce the correlations between the target and estimated Q values, thereby stabilizing the learning process of the DRL.

Compared with RL, DRL significantly reduces the model complexity, especially when it is applied to solve a variety of issues in the future communication systems. Additionally, DRL outperforms RL in terms of extracting system features to predict optimized rewards and actions. While RL only considers and updates the current reward and system state transition, DRL stores and samples historical rewards, actions, and state transitions into minibatches [3]. Then, DRL uses the stored historical data to train the deep network weight factors. Accordingly, the training process can be accelerated with the assistance of system and environment knowledge. The details

can be found in [4]. Moreover, cloud computing services and mobile edge computing devices can be integrated into forthcoming 5G and beyond (6G) networks. Mobile users and network operators will benefit from DRL by employing abundant computing capacities to solve resource management and network optimization. For example, the training process and historical data storage can be performed at the edge devices with GPU and massive memory space, respectively.

In the next section, we review a few recent works applying DRL to mobile 5G and beyond networks.

III. Deep Reinforcement Learning Applications for Future Generation Mobile Networks

In the literature, most of the optimization problems in future generation networks have been solved by centralized optimization approaches. However, these approaches make strong assumptions about complete information requirements on network conditions. As mobile networking environments become increasingly unpredictable, these assumptions turn impractical. Conversely, the DRL approach does not make the strong assumptions about the target system and hence becomes a practical technique for different issues arising in mobile 5G and beyond. In the following, we review some important works that capitalize DRL to address different issues in 5G networks.

A. Power Control and Power Management

Interference in 5G networks becomes more challenging to tackle because of an increasingly heterogeneous and dense environment such that conventional inter-cell interference coordination will become inefficient or even infeasible. A transmitter can decrease the transmit power to mitigate the interference. However, the data rate may be adversely affected. Interference mitigation is thus treated as the power control optimization problem of mobile networks. Power management of network devices, e.g., to turn on or turn off, is used to improve energy saving, hence reducing cost and carbon footprint. DRL allows the network entities to build knowledge about the networks and make its optimal decision towards power control and power management.

1) Power Control in Cellular Networks: Considering the uncertainty of dynamic 5G system, the authors in [5] developed a distributed dynamic power allocation scheme using DRL. In the scheme, each transmitter serves as an agent, and all agents are synchronized and take their actions simultaneously. Prior to taking actions, the agent is able to observe the effects resulted from the

past actions of the neighbors on the current decision period, based on which, the impacts of the current actions on future behaviors of the neighbors can be estimated. By using DRL, each agent determines a policy that maximizes the discounted expected future reward. Unlike conventional centralized optimization schemes, the computational complexity of the proposed DRL scheme does not depend on the network size, achieving high scalability when it is applied to the 5G networks with large coverage areas.

2) Power Management in Ultra Dense Networks: Ultra Dense Networks (UDN) suffers from high power consumption because of the dense deployment of Small Base Stations (SBSs). Thus, turning SBSs off dynamically is an effective solution to enhance energy efficiency. In [6], the authors formulated the SBSs ON/OFF problem in energy harvesting UDN into a dynamic optimization problem. Considering the uncertainty of energy charging, CSI, and traffic arrivals, DRL is introduced to learn a policy to deicide ON/OFF modes of SBSs for enhancing the energy efficiency. It is shown that the DRL-based ON/OFF scheduling scheme achieves higher energy efficiency than that achieved by Q Learning.

3) Power Control in mmWave Communications: With the deployment of mmWave communications in 5G, the Non-Line-of-Sight (NLOS) transmission is a critical issue. The authors in [7] proposed a dynamic transmission power control scheme to improve NLOS transmission performance. The objective is to maximize the total data rate achieved by all UEs in the 5G network under the constraints of transmission power and QoS requirements. Firstly, convolutionary neural networks is trained offline for estimating the Q-function. Then, DRL is executed online for UE association and power allocation actions.

B. Computation Offloading and Edge Caching

As computing becomes more important to support emerging mobile applications and services, the mobile 5G and beyond will be designed by deploying both computational resources and caching capabilities at the edge of mobile networks. This can significantly improve energy efficiency and QoS for applications that require intensive computations and low latency. The studies on computation offloading and edge caching in such a dynamic system typically involve complicated system analysis because of inherent couplings among heterogeneous users, QoS provisioning, mobility pattern, and radio resources. As such, a learning-based approach such as DRL becomes a viable solution to manage huge state space and optimization variables.

1) Mobile Edge Computation Offloading: In [8], the authors considered a UDN with a single User Equipment (UE) and multiple BSs. Computing tasks arrived from the UE can be offloaded to one of the BSs depending on channel qualities. Therein, the optimal offloading problem is modeled as an MDP, where the objective is to minimize the long-term expected cost. Considering the dynamics of time-varying channel qualities and the uncertainty of task arrivals, the DRL-based online strategic computational offloading algorithm was proposed to solve the formulated MDP problem. The numerical results show that the proposed algorithm leads to a significant improvement in terms of average cost. Different from [8], the authors in [9] discussed an edge computing system with a single edge server, where multiple UEs can perform computation offloading via wireless channels to the server. The sum cost of delay and energy consumption of all UEs is considered to be the optimization objective. To obtain the optimal policy with the curse of dimensionality avoided, the DRL-based approach is proposed to solve the computation offloading problem.

2) Edge Caching: Edge caching has the potential of reducing transmission cost, shortening latency, and relieving traffic loads of backhaul links. However, content caching involves the problem of policy control. In the problem, the network edge device with cache needs to decide contents to be stored in the cache under the circumstance that the content popularity distribution is always changing. In [10], the authors discussed a single BS as the cache node serving multiple mobile users, who keep requesting contents from the BS. To handle the requests efficiently, the authors utilized DRL for cache replacement decisions of the BS. The objective is to maximize the long-term cache hit rate, which is empirically confirmed by the numerical results.

3) Joint Edge Computing and Caching: Unlike the above works, which studied the edge computing and caching issues separately, the authors in [12] proposed an integrated framework that enables the orchestration of computing, networking, and caching resource management in vehicular networks. As the channel conditions of the BS, the computation capability of the edge servers and states of the cache nodes are all dynamically changing, the system state spaces are very large, and hence it is difficult to make a decision on which resources should be assigned to a specific vehicle. As such, they applied DRL to learn the best resource allocation policy. Likewise, the authors in [13] studied the joint computing, caching, and communication design problem for enhancing the performance of vehicular networks. The vehicle's mobility and the hard deadline delay are taken into account. DRL works by allowing the system to learn to minimize the system cost through exploration and exploitation of available actions.

C. Intelligent Transportation

The advance of 5G technology is envisioned to empower the real-time intelligent services offered by vehicular networks. As an important component of the development of the intelligent transportation systems, vehicular networks are expected to employ advanced communications and data collecting techniques to improve the design of reliable and efficient transportation systems. In [1], the authors developed a decentralized resource allocation scheme for vehicular networks based on DRL. The scheme is used to find the mapping between the partial observations of each vehicle and the optimal resource allocation solution. Specifically, this scheme can meet strict latency requirement on vehicle-to-vehicle links without requiring the accurate prior information.

D. Network Slicing

With advanced softwarization implemented in 5G networks, network slicing is treated as the key paradigm to fulfill the diversified service requirements [11]. The complexity of network resource management drives the need for slicing in mobile 5G networks, specifically, the exponential growth of wireless data service, diverse service requirements and heterogeneous wireless environments. As the name implies, slicing separates the network into different parts (slices), each of which is designed and optimized to meet specific service requirements. Network slicing is currently attracting tremendous interest from both academia and industry. In [14], the authors conducted the economic analysis for allocating requests of network slices, considering the maximum revenue of 5G infrastructure operator. The authors in [15] proposed an efficient resource slicing scheme for dynamically adjusting slice parameters to reduce the overall cost while maintaining the quality of service. Nevertheless, most of the slicing issues in 5G have been solved by constrained optimization [14], or distributed game-theoretic approaches [15]. As aforementioned, these approaches either make the strong assumptions about the objective functions (e.g., concavity) and data distribution (e.g., uniform distribution), or suffer from high time- and space-complexity. Since mobile 5G networks are becoming increasingly complex, the assumptions are unlikely to hold. Therefore, we resort to the DRL-based approach that does not make the strong assumptions about the system. It employs function approximation, which explicitly addresses the decision-making problems with large action and state spaces in 5G, such as network slicing. In network slicing, the diverse service requests from users in different slices are random and unpredictable. Thus, how to deal with the uncertain dynamics of service requests and huge action space of users is a crucial issue. However, the above problem has

not been well-addressed in existing literature. This motivates us to introduce the DRL-based approach to improve intelligent network slicing for mobile 5G networks.

IV. Deep Reinforcement Learning for Network Slicing in Mobile 5G

To address the resource and service optimization problems in 5G networks, we first present the future communication network with network slicing to provide differentiated services to mobile users. DRL is applied to the network to improve the efficiency of serving mobile user requests. Numerical results demonstrate that DRL can evidently improve the performance of the network slicing network.

A. DRL-Based Approach for Network Slicing

As a key feature of future generation communication networks, network slicing is introduced to virtualize network infrastructure to provide highly accessible resource and services to mobile users. However, the high diversity and complexity of the network architecture still raise challenges in resource and service allocations. Therefore, we propose the DRL-based scheme for network slicing optimization, which controls the allocation of user requests for network slices.

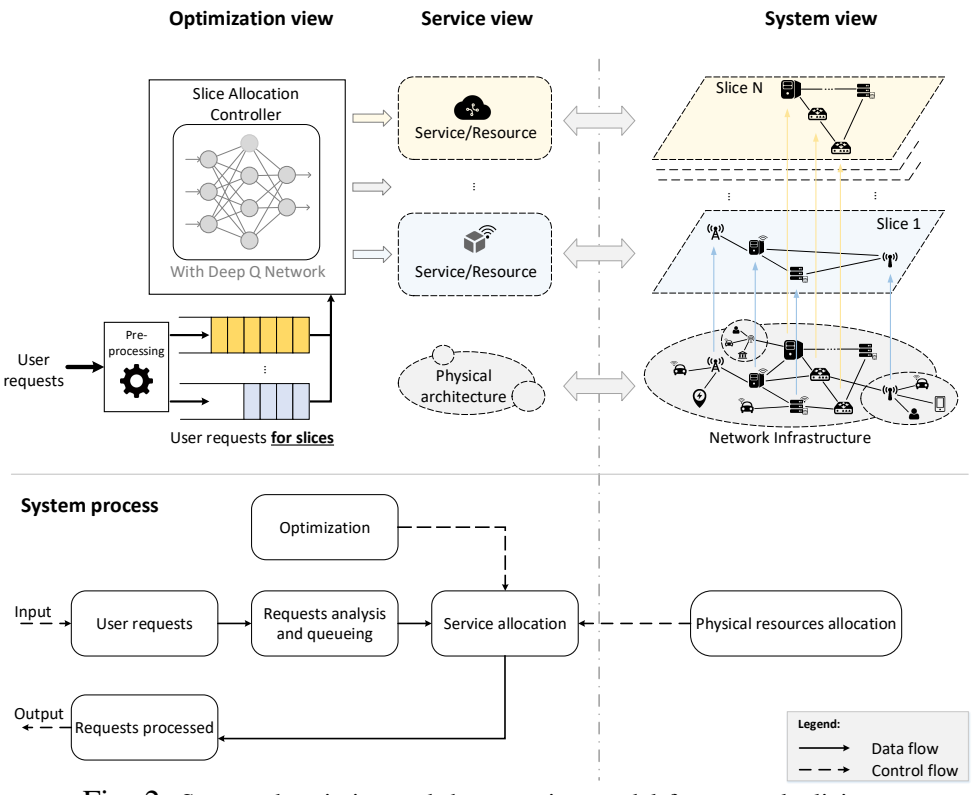


Fig. 2: System description and the queueing model for network slicing.

As shown in Figure 2, from the "system view", the physical infrastructure of a mobile system, i.e., system resources, is virtualized into multiple service slices. Each slice provides a set of services and network resources to serve incoming user requests, as shown by the "service view". With the service slice model, the users can request for different services provided by the network without being aware of the underlying physical infrastructure. However, the user requests for different services need to be scheduled by a controller to avoid conflicts of resource access. The scheduling process is depicted by the "optimization view", where the user request arrival process is represented as a queue model. Each incoming user request will be pre-processed and transformed into several slice requests, aiming for the corresponding network resources to be allocated to the slice. The slice requests are pushed into the queues based on their types. For example, a user request for cloud data storage services can be transformed into the slice requests for communications, i.e., wireless channel and Internet routing resources, and data persistence slice request, i.e., cloud storage. Here, the slice requests can be associated with different QoS requirements, e.g., service priorities and delay tolerance levels. The slice allocation controller retrieves the slice requests from the queues in the first-in-first-out fashion and allocates corresponding network slices with required network resources. Once the slice request is processed and served, i.e., allocated to an available slice, the slice allocation controller will remove the slice request from the queue. However, if there are insufficient network slices or network resources, the user requests will not be served and will be kept in the queues. The aforementioned steps for processing user requests is also summarized in Figure 2.

We proposed the DRL-based scheme for the slice allocation controller to optimize the slice request allocation, i.e., to allocate slice requests to the network slices, The DRL-based scheme observes the system states including the current levels of queue lengths and the currently available amount of resources in different slices. Once a slice request is successfully served, an immediate reward will be given to the slice allocation controller. The controller thus aims to maximize the utility gained, which can be defined as the reward from successfully serving the requests, minus the cost of the service delay, i.e., the queuing delay.

To achieve the utility maximization objective, the slice allocation controller implements a Deep Q Network (DQN) as a DRL-based approach to calculate the optimal slice allocation policies, as shown in Figure 2. The DQN implemented in the slice allocation controller is model-free, which does not require any prior information of the network. When the system is in operation, the controller keeps a set of sampled records of historical system states transitions,

and slice allocation policies. Based on the historical information, the controller can adjust the slice allocation policies dynamically, which is defined as a training process.

In the follows, we evaluate the performance of the proposed DRL-based scheme.

B. Numerical Results

1) Parameter Setting: We evaluate the performance of the proposed DRL-based scheme for network slicing optimization with the following network setting and parameters:

- The network supports two types of services, i.e., guaranteed service and best effort service.[1] The former has a higher priority, i.e., generating more reward if its request is served, than that of the latter;
- A guaranteed service queue (GSQ) with the maximum capacity of 200 requests is to store user requests with the guaranteed service slices, e.g., video stream data;
- A best-effort service queue (BeSQ) with the maximum capacity of 200 requests is to store user requests with best-effort service slices, e.g., delay tolerant data;
- There are three types of network resources, e.g., computing, storage, and wireless bandwidth;
- The action indicates the numbers of slice requests retrieved from GSQ and BeSQ to serve.

The request arrival rates for the GSQ and the BeSQ are both 0.35, and the successfully processing rates are 0.35 and 0.95, respectively. The successful processing rates indicate that the slice requests from the GSQ are more difficult for the network to fulfill their requirements, for example, because of more network resource required and strict service guarantee. Immediate rewards for successfully processing slice requests from the GSQ and the BeSQ are set to be 50 and 10, respectively. The delay costs of slice requests waiting in the GSQ and the BeSQ are 0.05 and 0.01 per request per time slot, respectively.

We consider baseline schemes, to compare with the proposed DRL-based scheme, including (i) a conventional Q Learning with a table to store the mapping of system state-action pairs to Q-values, (ii) a greedy scheme only maximizing the immediate rewards obtained from taking current actions, and (iii) a scheme taking random actions.

2) Simulation: In the simulation, the DRL-based scheme and baseline schemes are executed for 800 episodes.

[1]Only two services are considered to ease the interpretation and presentation of the results. Nevertheless, the model can be easily extended for any number of service types.

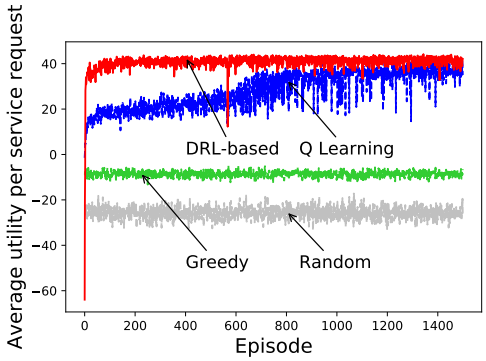
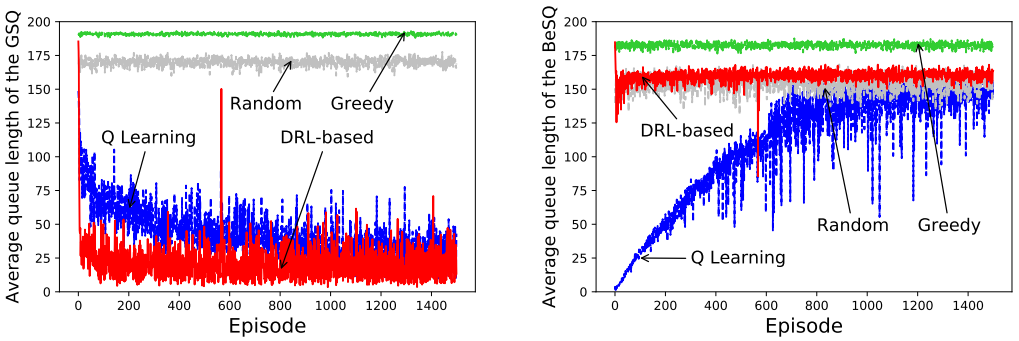


Fig. 3: Performance evaluation of the DRL-based scheme: Average gained utility per service.

Figure 3 shows the average utility gained from processing and serving a slice request. The DRL-based scheme outperforms the other baseline schemes. During the initial episodes, i.e., episode 0 to episode 50, the DRL-based scheme generates a low average utility because of the untrained DQN. After about 150 episodes, the average utility obtained by the DRL-based scheme reaches the maximum and becomes stable. The reason is that system state transitions and utilities, i.e., immediate rewards minus costs, of the previous episodes are sampled to train the DQN. The DRL-based scheme uses the historical data to adjust the weights in the DQN to approximate the optimal action outputs more accurately.



(a) (b)

Fig. 4: (a) Average length of Guaranteed Service Queue (GSQ) and (b) average length of Best-effort Service Queue (BeSQ).

Figures 4(a) and (b) show the average queue lengths of the GSQ and the BeSQ, respectively. From the figures, both DRL-based and Q Learning schemes can efficiently process slice requests. In Figure 4(a), the DRL-based scheme has a decreasing average queue length for the GSQ, which is also lower than that of the Q Learning scheme. This indicates that the DRL-based scheme can learn from historical data that keeping slice requests in the GSQ for long has resulted in the high delay cost as well as the small reward. Consequently, the DRL-based scheme quickly adjusts its policy, through the learning process, to serve more slice requests from the GSQ to improve the utility. In Figure 4(b), the average queue length of the BeSQ increases when the

DRL-based and the Q Learning schemes are adopted. The reason is that the DRL-based scheme learns to give a higher priority to the slice requests from the GSQ than those from the BeSQ. By contrast, Q Learning is relatively ineffective in doing so. This is demonstrated by the results of the DRL-based and the Q Learning schemes in Figure 4(b). However, after the training, the slice allocation controller prefers to serve the slice requests from the GSQ, as shown in Figure 4(a). As a result, the average queue length of the BeSQ increases.

V. Conclusion

In this article, we have presented fundamental concepts of Deep Reinforcement Learning (DRL) to deal with decision making, resource and service allocation problems in future generation communication networks, i.e., 5G and beyond. We have reviewed the recent works applying DRL in 5G networks. As an example of the DRL application, we have proposed the DRL-based scheme for network slicing optimization. We have performed numerical studies to show the capability and optimized performance metrics of the DRL-based scheme.

Nevertheless, the existing works only scratch the surface, and the potential of DRL to tackle many more problems in 5G networks and beyond (6G) is yet to be explored. We outline some future research directions.

- *Multi-agent DRL in 5G:* At the dawn of 5G, we foresee an enormous increase in the number of pervasively connected Internet of Things (IoT) devices, and most of the IoT sensors are owned by device owners instead of operators. Thus, the traditional approach of the customizations of DRL for an individual network entity does not work well in a heterogeneous 5G-enabled IoT which consists of multiple stakeholders with fast-changing network conditions. The interactions among multiple IoT device owners, i.e., the agents, further complicate the network resource and service management in which considerable increases in the state and action spaces of the problem are expected. This situation inevitably slows down the learning algorithms and compromises the performance of the learning policy.
- *Distributed DRL in 5G:* The DRL requires the training of DQNs, which is implemented at a centralized network controller with large computational capacity. However, the centralized DRL does not match well with the distributed 5G systems including massive IoT devices. As a result, it is necessary to design distributed implementation of DRL that decomposes resource-intensive DQN training task into different sub-tasks for individual devices with moderate or limited computing power.

REFERENCES

[1] H. Ye, L. Liang, G. Y. Li, J. Kim, L. Lu and M. Wu, "Machine learning for vehicular networks: Recent advances and application examples," *IEEE Vehicular Technology Magazine*, vol. 13, no. 2, pp. 94–101, 2018.

[2] L. P. Kaelbling, M. L. Littman and A. W. Moore, "Reinforcement learning: A survey," *Journal of artificial intelligence research*, vol. 4, pp. 237–285, 1996.

[3] V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529, 2015.

[4] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang and D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," *arXiv preprint arXiv:1810.07862*, 2018.

[5] Y. S. Nasir and D. Guo, "Deep reinforcement learning for distributed dynamic power allocation in wireless networks," *arXiv preprint arXiv:1808.00490*, 2018.

[6] H. Li, H. Gao, T. Lv and Y. Lu, "Deep Q learning based dynamic resource allocation for self-powered ultra-dense networks," in *Proceedings of IEEE International Conference on Communications Workshops*, Kansas City, MO, May 2018.

[7] C. Luo, J. Ji, Q. Wang, L. Yu and P. Li, "Online power control for 5G wireless communications: A deep q-network approach," in *Proceedings of IEEE International Conference on Communications*, Kansas City, MO, May 2018.

[8] X. Chen, H. Zhang, C. Wu, S. Mao, Y. Ji and M. Bennis, "Optimized computation offloading performance in virtual edge computing systems via deep reinforcement learning," *IEEE Internet of Things Journal*, to appear.

[9] J. Li, H. Gao, T. Lv and Y. Lu, "Deep reinforcement learning based computation offloading and resource allocation for MEC," in *Proceedings of IEEE Wireless Communications and Networking*, Barcelona, Spain, April 2018.

[10] C. Zhong, M. C. Gursoy and S. Velipasalar, "A deep reinforcement learning-based framework for content caching," in *Annual Conference on Information Sciences and Systems*, Princeton, NJ, March 2018.

[11] X. Foukas, G. Patounas, A. Elmokashfi and M. K. Marina, "Network slicing in 5G: Survey and challenges," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 94–100, 2017.

[12] Y. He, N. Zhao and H. Yin, "Integrated networking, caching, and computing for connected vehicles: A deep reinforcement learning approach," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 1, pp. 44–55, 2018.

[13] L. T. Tan and R. Q. Hu, "Mobility-aware edge caching and computing in vehicle networks: A deep reinforcement learning," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 11, pp. 10190–10203, 2018.

[14] D. Bega, M. Gramaglia, A. Banchs, V. Sciancalepore, K. Samdanis and X. Costa-Perez, "Optimising 5g infrastructure markets: The business of network slicing," in *IEEE Conference on Computer Communications*, Atlanta, GA, May 2017.

[15] P. Caballero, A. Banchs, G. de Veciana, X. Costa-Pérez and A. Azcorra, "Network slicing for guaranteed rate services: Admission control and resource allocation games," *IEEE Transactions on Wireless Communications*, vol. 17, no. 10, pp. 6419–6432, 2018.